分布式系统 Lab2 文档

1.数据预处理

1.1 结果展示

运行 createFiles.py，即可在当前目录下面生成名字为 Files 的文件夹，文件夹中的每个文件存放着标题和内容。

#（关注夏收）（3）夫妻"麦客"忙麦收.txt	2023/12/7 14:02	文本文档	2 KB
（文化）（2）上海老相机制造博物馆即将开门...	2023/12/7 14:02	文本文档	1 KB
"韩影宫"集体关门　六千多会员被"套"（.txt	2023/12/7 14:02	文本文档	1 KB
"十分钟"传导"三公里".txt	2023/12/7 14:02	文本文档	5 KB
"在野党"联手　台湾行政负责人施政报告受阻.txt	2023/12/7 14:02	文本文档	2 KB
"走婚族"悄然兴起　六成网友称能接受.txt	2023/12/7 14:02	文本文档	3 KB
·搜狐社区用户请确认您发表的言论符合.txt	2023/12/7 14:02	文本文档	1 KB
5月数据或令调控再出手　非对称降息预期升温...	2023/12/7 14:02	文本文档	4 KB
１０万七天理财收益仅百元　银行产品实际收益...	2023/12/7 14:02	文本文档	5 KB
１７岁癌症少年无钱治疗　含泪捐全身器官（.txt	2023/12/7 14:02	文本文档	1 KB
２０１０哪些"浮云"让你印象深刻.txt	2023/12/7 14:02	文本文档	1 KB
A股市场遭遇黑色星期一　基金表态不必过度悲...	2023/12/7 14:02	文本文档	3 KB
ｈｏｌｄ不住命运的中国人.txt	2023/12/7 14:02	文本文档	1 KB
巴基斯坦惊现真实版"象人"（.txt	2023/12/7 14:02	文本文档	1 KB
北京８家银行房贷利率最低８．５折　业内称已...	2023/12/7 14:02	文本文档	4 KB
北京宝马女挤翻本田撞飞路人　受审迟到庭上轻...	2023/12/7 14:02	文本文档	1 KB
北新建材及其控股子公司拟逾４亿投资三项目.txt	2023/12/7 14:02	文本文档	2 KB
背景资料：印度概况.txt	2023/12/7 14:02	文本文档	1 KB
财政部：政府性资金要对民间投资主体同等对待...	2023/12/7 14:02	文本文档	3 KB
拆迁款瘦身记：拨款１７１万支付１４０万到手...	2023/12/7 14:02	文本文档	2 KB
超惊险！蟒蛇暴起咬人全过程！.txt	2023/12/7 14:02	文本文档	1 KB

1.2 代码设计思路

首先读取 document.dat 文件（使用 encoding='gb18030' ） 。

然后通过正则表达式提取标题 <contenttitle> 标签和内容 <content> 标签内的内容。

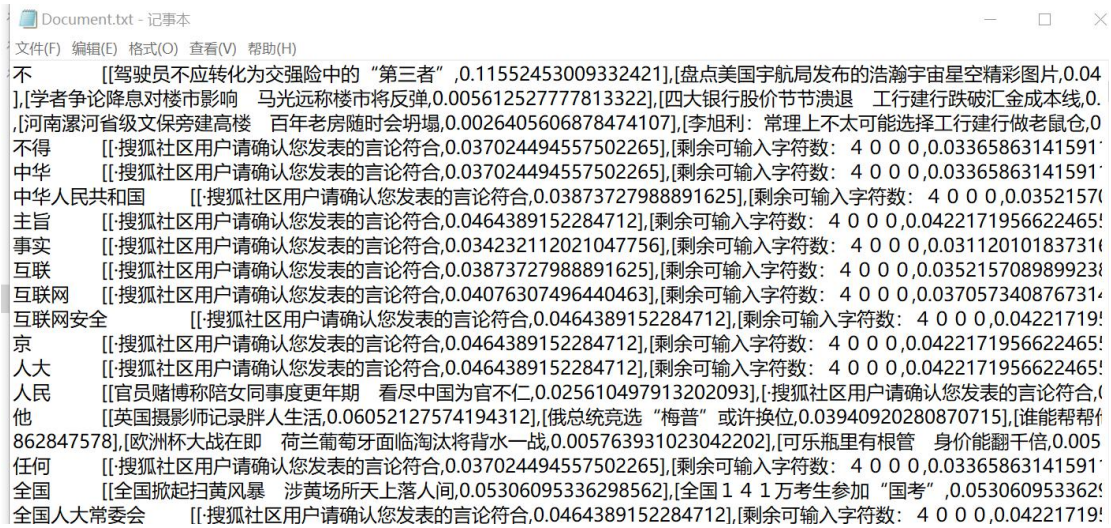
通过对标题+内容使用 jieba 库分词，再将每一个 title 和 content 的内容写入一个 txt 文件即可。

2.TF-IDF 文档生成

2.1 结果展示

运行 createtfidf.py，并且在 Pycharm 的 Edit configuration 中加上参数./Files/*即可。

运行结果 Document.txt 如下图所示：



2.2 代码实现思路

首先要计算得出一个文件中所有的 word 和该 word 出现了多少次，根据文档提供的 MapReduce 实例实现，从而计算出 TF。

接下来计算所有 word 的出现次数，计算出 IDF。

最后计算 TF-IDF，构建输出映射，对输出映射中的内容进行排序，并将结果写入 Document.txt 文件中。该文件每一行的格式如下：

word1 [file1, TF-IDF1], [file2, TF-IDF2], ...

2.3 函数说明

get_file_and_word_cnt_map(bytes_stream)：从字节流中获取文件名和词数，并返回一个字典。

calculate_tf(total_map, single_map)：计算每个文档中每个词的 TF 值。

calculate_idf(word_map, file_cnt)：计算逆文档频率（IDF）并更新每个单词的出现次数。

calculate_files(directory): 计算目录下的文件总数。

calculate_word_freq_in_file(word_cnt_in_map): 计算每个词在文档中出现的次数。

calculate_word_cnt_from_byte(bytes_stream): 从字节流中提取每个文档中单词的计数信息，并返回一个嵌套字典。

calculate_tf_idf(TF_Map, IDF_Map): 计算 TF-IDF。

create_map(word_map, tf_idf_map): 根据每个单词的出现次数和 TF-IDF 映射构建输出映射。

输出映射的格式如下：word1 [file1, TF-IDF1], [file2, TF-IDF2]

sort_map_res(map_res): 对输出映射中的内容的 TF-IDF 值进行从大到小的排序。

最后将排序完成的内容输出到 Document.txt 中。

3.文档查询

3.1 结果展示

运行 tfidf_search.py，然后输入关键词即可。

```
Please input keyword: 上海
上海山寨版五角大楼
台考试院长独生女　婚后１年坠楼身亡
（文化）（２）上海老相机制造博物馆即将开门迎客
新股申报脚步放缓　上周新增３家拟上市公司
扎克伯格携妻罗马当街吃３０元麦当劳午餐（组图）
三部委决定在上海试行启运港退税政策
各国央行推量化动力足　中国或面临两难格局
中美联手破获跨国走私武器案　主犯系美国士兵
证券业协会六招遏制ＩＰＯ人情报价
国内油价两连跌　下游行业跟涨容易跟跌难
公安机关销毁１０余万非法枪支　跨国武器走私渐起
```

3.2 代码实现思路

首先读入 Document.txt，通过正则表达式找到词、文件名、TF-IDF 值，将其存入字典中。接着提示用户输入 keyword，若存在则将文档序列实现从大到小的输出。